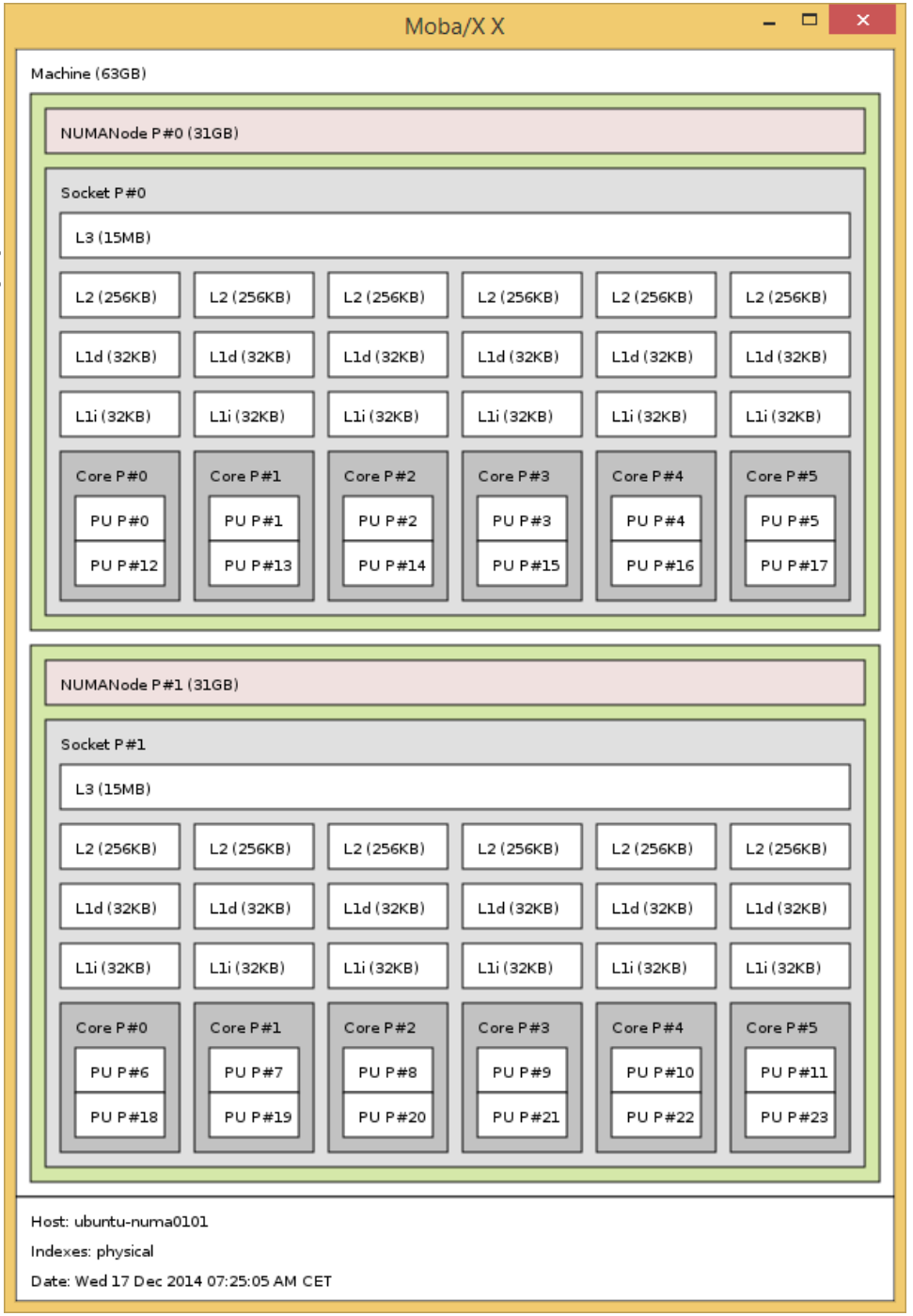
topology detection

- Screenshot


Moba/X X
Machine (63GB)
NUMANode P#0 (31GB)
Socket P#0
L3 (15MB)
L2 (256KB)
L1d (32KB)
L1i (32KB)
Core P#0
PU P#0
PU P#12
Core P#1
PU P#1
PU P#13
Core P#2
PU P#2
PU P#14
Core P#3
PU P#3
PU P#15
Core P#4
PU P#4
PU P#16
Core P#5
PU P#5
PU P#17
NUMANode P#1 (31GB)
Socket P#1
L3 (15MB)
L2 (256KB)
L1d (32KB)
L1i (32KB)
Core P#0
PU P#6
PU P#18
Core P#1
PU P#7
PU P#19
Core P#2
PU P#8
PU P#20
Core P#3
PU P#9
PU P#21
Core P#4
PU P#10
PU P#22
Core P#5
PU P#11
PU P#23
Host: ubuntu-numa0101
Indexes: physical
Date: Wed 17 Dec 2014 07:25:05 AM CET

/proc/cpuinfo

```
processor       : 0
vendor_id       : GenuineIntel
model name      : Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz

cache size      : 15360 KB
physical id     : 0
siblings        : 12
core id         : 0
cpu cores       : 6

flags           : fpu vme de pse …

processor       : 1
vendor_id       : GenuineIntel
model name      : Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz
cache size      : 15360 KB
physical id     : 0
siblings        : 12
core id         : 1
cpu cores       : 6
flags           : fpu vme de pse …
```

/proc/cpuinfo

```
processor       : 0
vendor_id       : GenuineIntel
model name      : Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz

cache size      : 15360 KB
physical id     : 0
siblings        : 12
core id         : 0
cpu cores       : 6

flags           : fpu vme de pse …

processor       : 1
vendor_id       : GenuineIntel
model name      : Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz
cache size      : 15360 KB
physical id     : 0
siblings        : 12
core id         : 1
cpu cores       : 6
flags           : fpu vme de pse …
```

numactl --hardware

```
node 0 cpus: 0 1 2 3 4 5 12 13 14 15 16 17
node 0 size: 32140 MB
node 0 free: 28841 MB
node 1 cpus: 6 7 8 9 10 11 18 19 20 21 22 23
node 1 size: 32251 MB
node 1 free: 30743 MB
node distances:
node   0   1
  0:  10  20
  1:  20  10
```

numactl --hardware

```
node 0 cpus: 0 1 2 3 4 5 12 13 14 15 16 17
node 0 size: 32140 MB
node 0 free: 28841 MB
node 1 cpus: 6 7 8 9 10 11 18 19 20 21 22 23
node 1 size: 32251 MB
node 1 free: 30743 MB
node distances:
node   0   1
  0:  10  20
  1:  20  10
```

numactl --hardware

```
[...]
node distances:
node   0   1   2   3   4   5   6   7
  0:  10  12  17  17  19  19  19  19
  1:  12  10  17  17  19  19  19  19
  2:  17  17  10  12  19  19  19  19
  3:  17  17  12  10  19  19  19  19
  4:  19  19  19  19  10  12  17  17
  5:  19  19  19  19  12  10  17  17
  6:  19  19  19  19  17  17  10  12
  7:  19  19  19  19  17  17  12  10
```

- /sys/devices/system/node/node*
- /sys/devices/system/node/node*/distance
- /sys/devices/system/node/node*/cpu*

Information sources

- ACPI
  - System Resource Affinity Table (SRAT)
    - NUMA nodes (proximity domains), RAM, Interrupt controllers,
  - System Locality Distance Information Table (SLIT)
    - only relative information about "distances"

estimate hardware topology

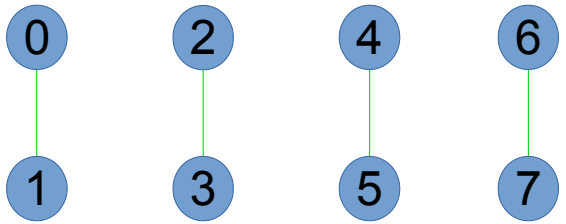
```
[...]
node distances:
node   0   1   2   3   4   5   6   7
  0:  10  12  17  17  19  19  19  19
  1:  12  10  17  17  19  19  19  19
  2:  17  17  10  12  19  19  19  19
  3:  17  17  12  10  19  19  19  19
  4:  19  19  19  19  10  12  17  17
  5:  19  19  19  19  12  10  17  17
  6:  19  19  19  19  17  17  10  12
  7:  19  19  19  19  17  17  12  10
```

estimate hardware topology

```
[...]
node distances:
node   0   1   2   3   4   5   6   7
  0:  10  12  17  17  19  19  19  19
  1:  12  10  17  17  19  19  19  19
  2:  17  17  10  12  19  19  19  19
  3:  17  17  12  10  19  19  19  19
  4:  19  19  19  19  10  12  17  17
  5:  19  19  19  19  12  10  17  17
  6:  19  19  19  19  17  17  10  12
  7:  19  19  19  19  17  17  12  10
```

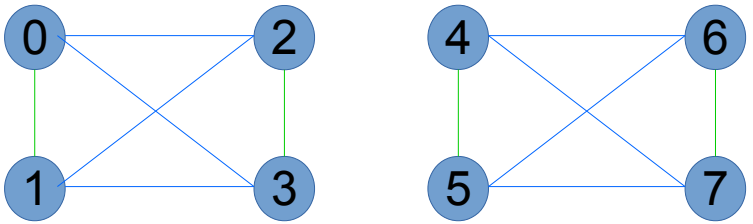
estimate hardware topology

```
[...]
node distances:
node   0   1   2   3   4   5   6   7
  0:  10  12  17  17  19  19  19  19
  1:  12  10  17  17  19  19  19  19
  2:  17  17  10  12  19  19  19  19
  3:  17  17  12  10  19  19  19  19
  4:  19  19  19  19  10  12  17  17
  5:  19  19  19  19  12  10  17  17
  6:  19  19  19  19  17  17  10  12
  7:  19  19  19  19  17  17  12  10
```



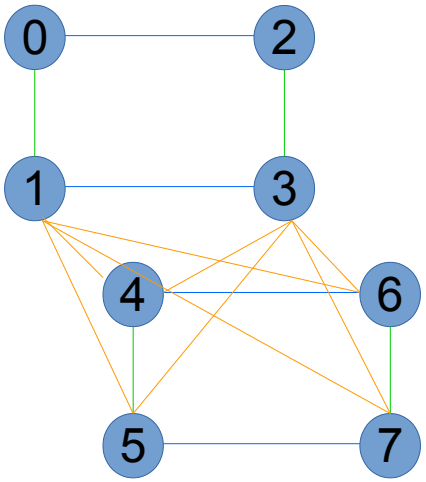
estimate hardware topology

```
[...]
node distances:
node   0   1   2   3   4   5   6   7
  0:  10  12  17  17  19  19  19  19
  1:  12  10  17  17  19  19  19  19
  2:  17  17  10  12  19  19  19  19
  3:  17  17  12  10  19  19  19  19
  4:  19  19  19  19  10  12  17  17
  5:  19  19  19  19  12  10  17  17
  6:  19  19  19  19  17  17  10  12
  7:  19  19  19  19  17  17  12  10
```



estimate hardware topology

```
[...]
node distances:
node   0   1   2   3   4   5   6   7
  0:  10  12  17  17  19  19  19  19
  1:  12  10  17  17  19  19  19  19
  2:  17  17  10  12  19  19  19  19
  3:  17  17  12  10  19  19  19  19
  4:  19  19  19  19  10  12  17  17
  5:  19  19  19  19  12  10  17  17
  6:  19  19  19  19  17  17  10  12
  7:  19  19  19  19  17  17  12  10
```



?

actual hardware topology?

- Wasn't there only max 4 QPI links per CPU?
- `$cat /sys/devices/virtual/dmi/id/product_name`

  `ProLiant DL980 G7`

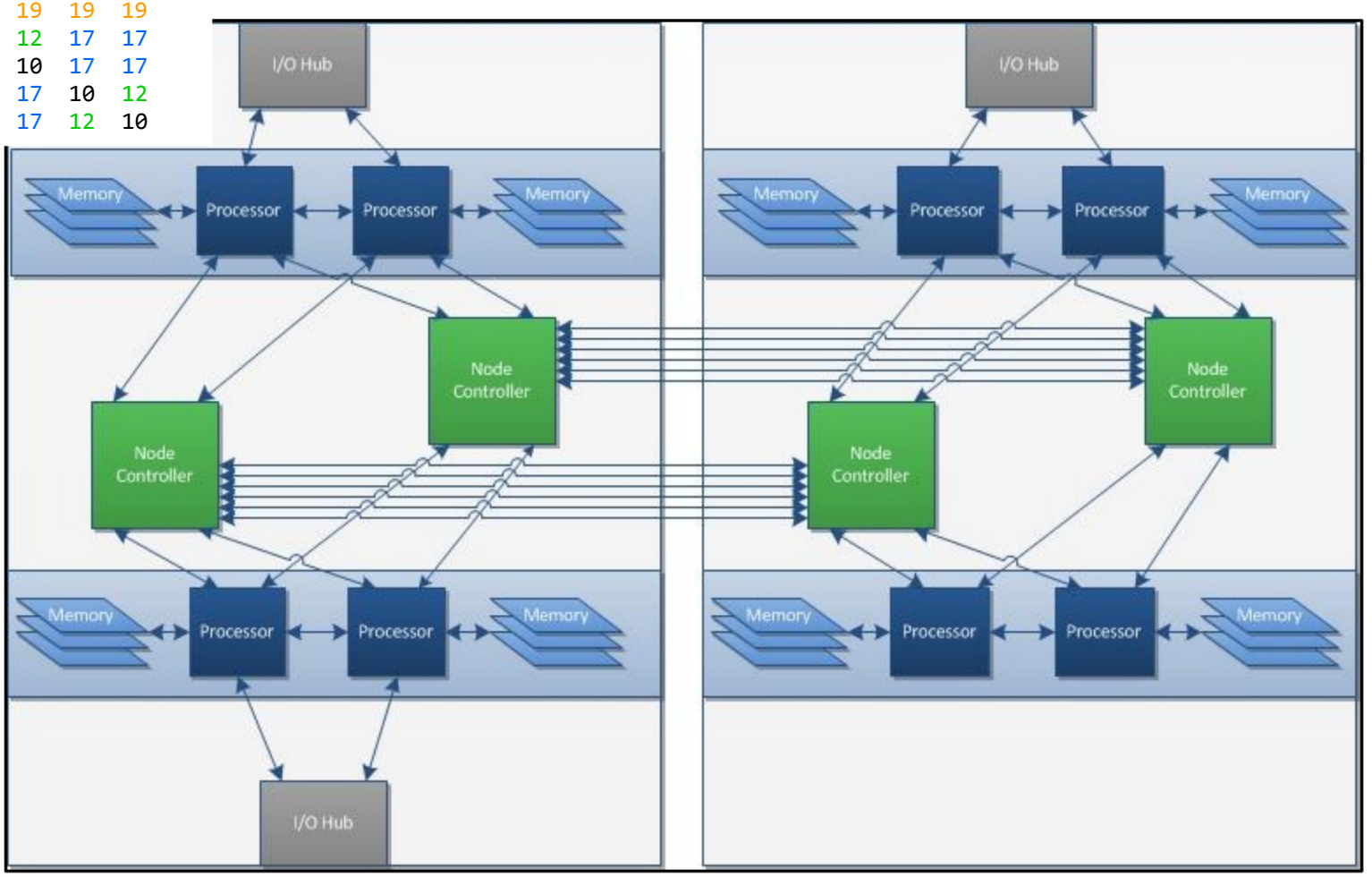
actual hardware topology?

- Wasn't there only max 4 QPI links per CPU?
- `$cat /sys/devices/virtual/dmi/id/product_name`

  `ProLiant DL980 G7`

actual hardware topology

```
[...]
node distances:
node   0   1   2   3   4   5   6   7
  0:  10  12  17  17  19  19  19  19
  1:  12  10  17  17  19  19  19  19
  2:  17  17  10  12  19  19  19  19
  3:  17  17  12  10  19  19  19  19
  4:  19  19  19  19  10  12  17  17
  5:  19  19  19  19  12  10  17  17
  6:  19  19  19  19  17  17  10  12
  7:  19  19  19  19  17  17  12  10
```



Absolute measurements

- `mlc -e`

Absolute measurements

Raw Bandwidth values (MB/s)

Socket	0	1	2	3	4	5	6	7
0	25368,9	10882,8	11203,2	11201,7	11179,1	11175,9	11175,6	11177,8
1	10858,8	25524,8	11184,3	11186,9	11168,0	11183,7	11175,1	11165,0
2	11165,7	11211,5	25605,7	10877,3	11182,6	11173,5	11175,9	11182,9
3	11181,9	11207,4	10876,2	25558,8	11176,5	11167,7	11169,2	11174,5
4	11119,3	11166,1	11165,6	11169,0	25540,0	10862,3	11186,2	11190,0
5	11140,0	11179,1	11182,0	11172,1	10867,8	25553,2	11187,6	11192,8
6	11153,3	11191,1	11193,1	11189,3	11212,0	11208,4	25633,1	10900,8
7	11165,8	11196,0	11195,1	11187,2	11210,4	11210,6	10920,2	25634,7

Absolute measurements

Raw Bandwidth values (MB/s) → non-direct connections very similar speeds

Socket	0	1	2	3	4	5	6	7
0	25368,9	10882,8	11203,2	11201,7	11179,1	11175,9	11175,6	11177,8
1	10858,8	25524,8	11184,3	11186,9	11168,0	11183,7	11175,1	11165,0
2	11165,7	11211,5	25605,7	10877,3	11182,6	11173,5	11175,9	11182,9
3	11181,9	11207,4	10876,2	25558,8	11176,5	11167,7	11169,2	11174,5
4	11119,3	11166,1	11165,6	11169,0	25540,0	10862,3	11186,2	11190,0
5	11140,0	11179,1	11182,0	11172,1	10867,8	25553,2	11187,6	11192,8
6	11153,3	11191,1	11193,1	11189,3	11212,0	11208,4	25633,1	10900,8
7	11165,8	11196,0	11195,1	11187,2	11210,4	11210,6	10920,2	25634,7

Absolute measurements

Idle latency

Soc	0	1	2	3	4	5	6	7
0	37.30	49.90	74.70	73.50	84.40	82.60	80.30	82.00
1	49.40	36.30	73.80	76.90	85.00	82.30	84.40	86.10
2	75.00	75.60	35.00	46.40	75.30	75.40	78.10	81.80
3	69.60	67.60	47.30	35.10	83.50	81.70	78.20	81.50
4	77.10	78.60	77.50	78.90	34.80	48.90	70.00	75.30
5	79.80	76.50	79.80	80.90	46.20	35.10	69.20	69.80
6	75.70	74.10	80.00	77.10	67.60	67.80	35.10	46.50
7	83.70	84.00	84.00	82.70	69.90	70.30	47.40	34.90

Normalisiert: SLIT (expected) vs measured

10	10
12	11,1
17	22,7
19	25,5

Absolute measurements

Idle latency

Soc	0	1	2	3	4	5	6	7
0	37.30	49.90	74.70	73.50	84.40	82.60	80.30	82.00
1	49.40	36.30	73.80	76.90	85.00	82.30	84.40	86.10
2	75.00	75.60	35.00	46.40	75.30	75.40	78.10	81.80
3	69.60	67.60	47.30	35.10	83.50	81.70	78.20	81.50
4	77.10	78.60	77.50	78.90	34.80	48.90	70.00	75.30
5	79.80	76.50	79.80	80.90	46.20	35.10	69.20	69.80
6	75.70	74.10	80.00	77.10	67.60	67.80	35.10	46.50
7	83.70	84.00	84.00	82.70	69.90	70.30	47.40	34.90

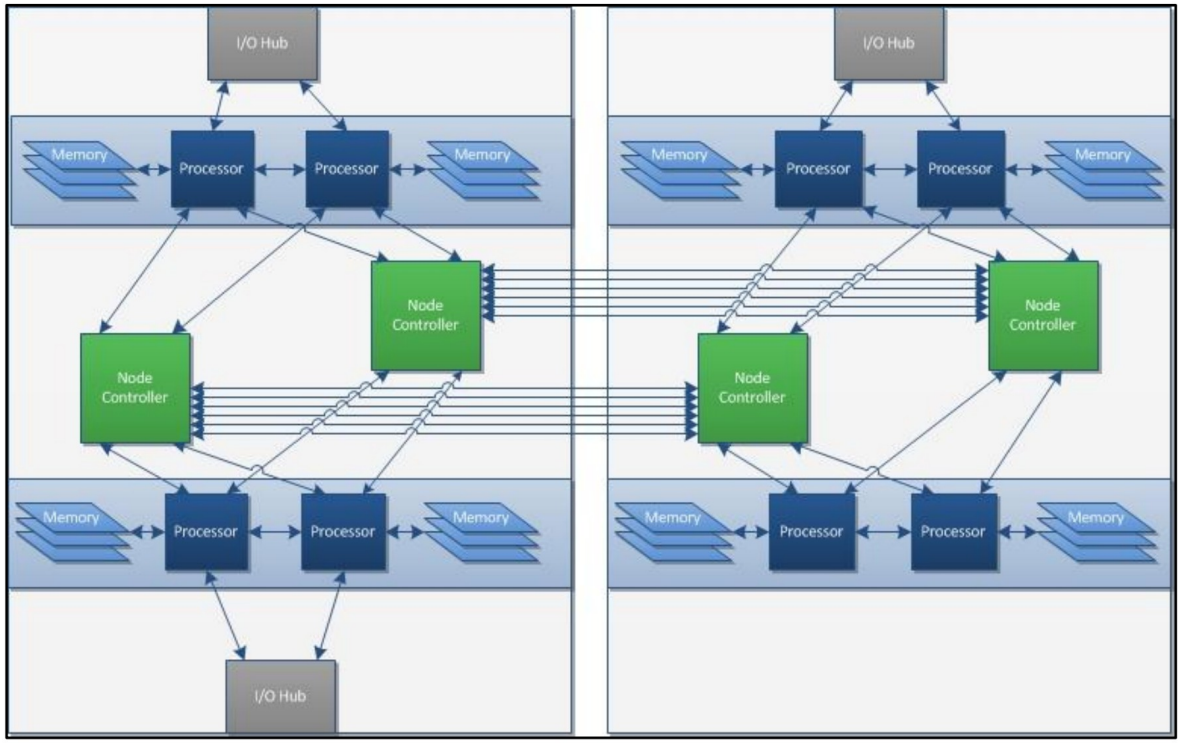
SLIT (expected) vs measured (fastest normalized to 10)

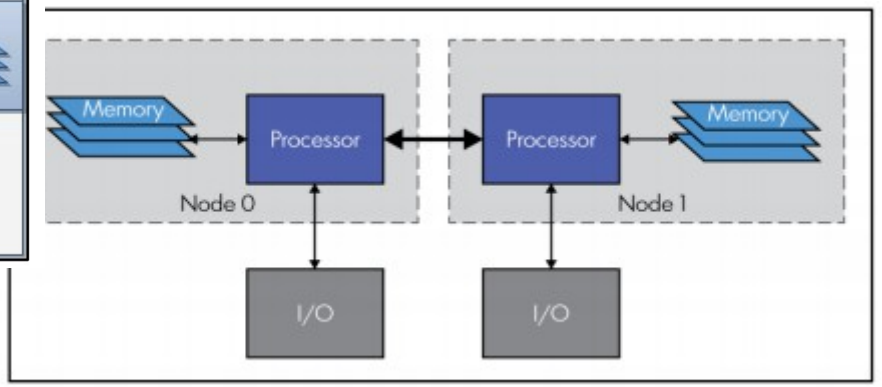
10	10
12	11,1
17	22,7
19	25,5

Close...ish?

NUMA and IO

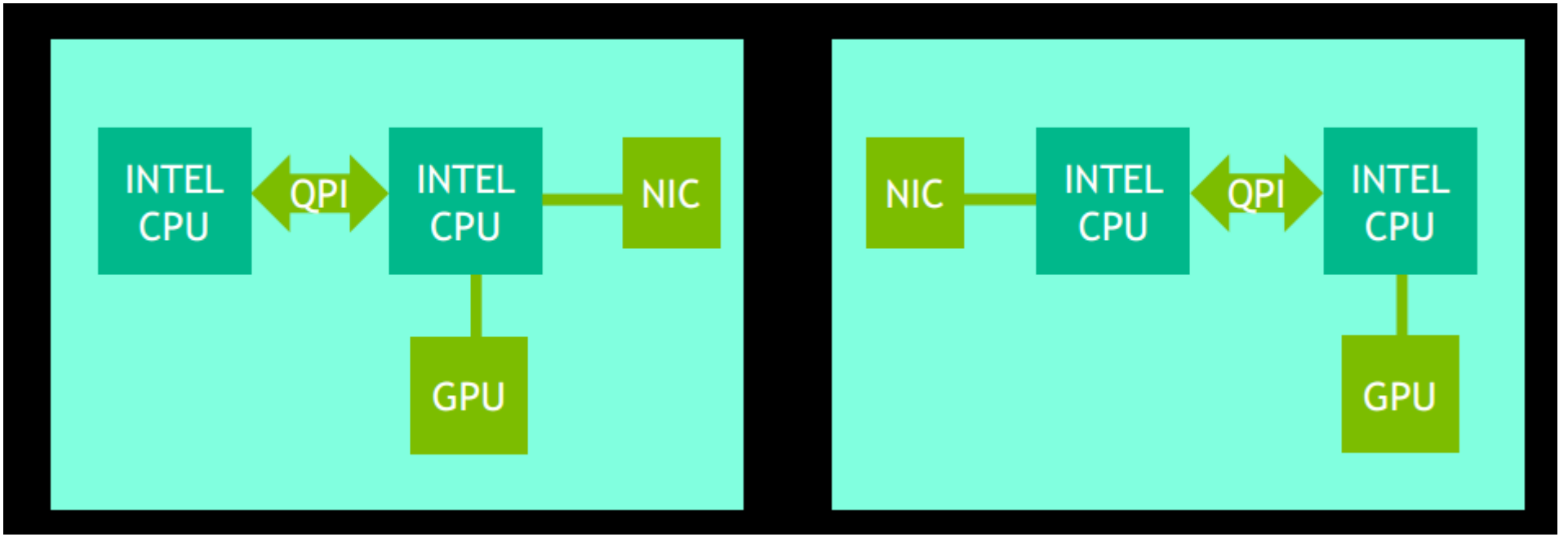
- not only memory is non-uniform
- IO controllers are connected to only some nodes

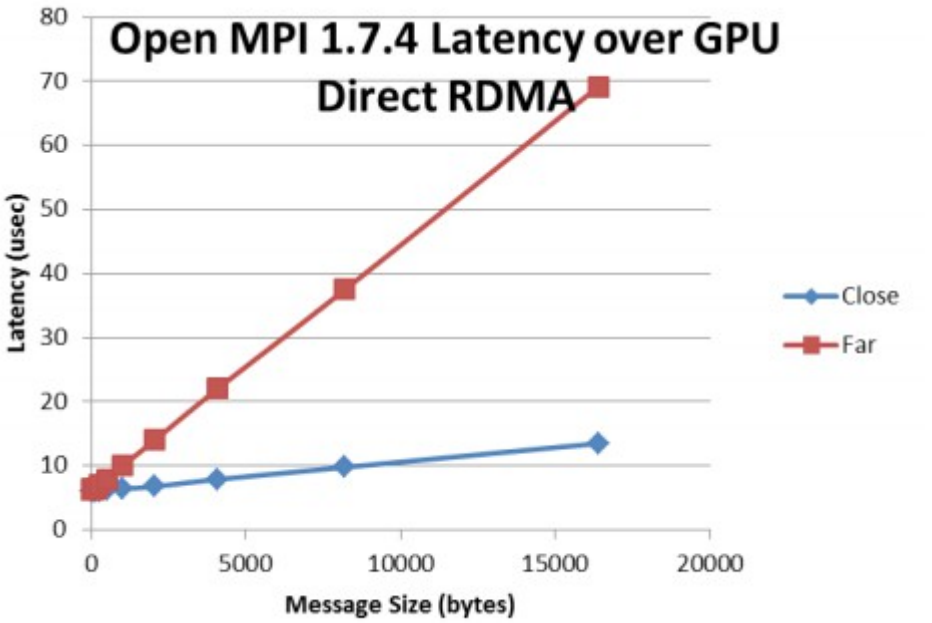


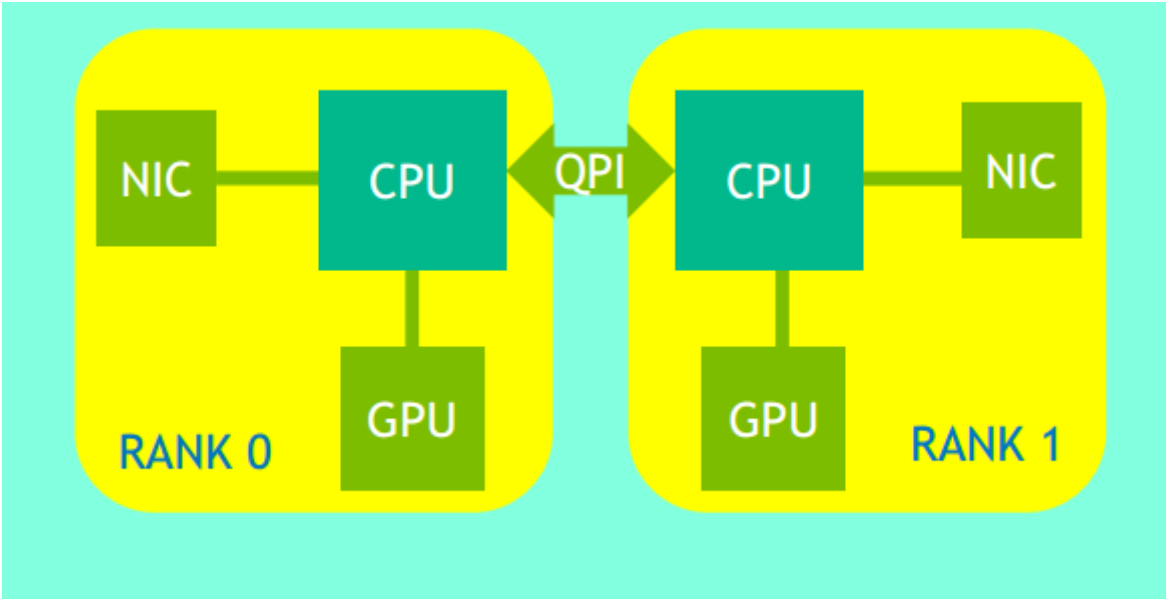


NUMA and IO

- PCIe 4.0: 2 GB/s per lane vs QPI: 12,8 GB/s per link
- 40 GbE cards, access to GPU memory utilize these speeds
- example applications: network processing, clusters


INTEL CPU
QPI
INTEL CPU
NIC
GPU
NIC
INTEL CPU
QPI
INTEL CPU
GPU


Open MPI 1.7.4 Latency over GPU Direct RDMA
Latency (usec)
Message Size (bytes)
0
10
20
30
40
50
60
70
80
0
5000
10000
15000
20000
Close
Far


NIC
CPU
QPI
CPU
NIC
GPU
GPU
RANK 0
RANK 1

Identifying IO nodes

- lspci -tv, then look up in /sys/bus/...
- Why is our machine structured like it is?

future work

- do extended measurements on 8-node machine
- utilize QPI link information for discovery

- http://www.acpi.info/DOWNLOADS/ACPIspec40a.pdf
- http://h50146.www5.hp.com/products/software/oe/linux/mainstream/support/whitepaper/pdfs/c03261871.pdf
- http://on-demand.gputechconf.com/gtc/2014/presentations/S4589-openmpi-rdma-support-cuda.pdf
- man numa, numactl, lstopo
- http://www.ntop.org/pf_ring/not-all-servers-are-alike-with-pf_ring-zcdna-part-3/